Week14_예습과제_김도희

Flamingo: a Visual Language Model for Few-Shot Learning

Flamingo : few-shot으로 다양한 task를 빠르게 적응 및 수행할 수 있는 VLM

1. 사전 학습된 비전 전용 및 언어 전용 모델을 연결하는 방법.
2. 텍스트와 이미지를 임의로 섞은 시퀀스를 처리할 수 있는 능력.
3. 이미지나 영상을 입력으로 원활하게 처리할 수 있는 구조.

1. Introduction

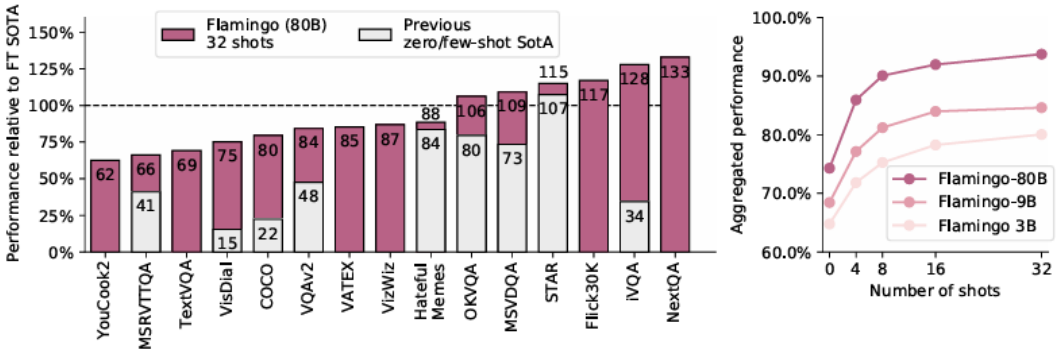

Figure 2: **Flamingo results overview.** *Left*: Our largest model, dubbed *Flamingo*, outperforms state-of-the-art fine-tuned models on 6 of the 16 tasks we consider with no fine-tuning. For the 9 tasks with published few-shot results, *Flamingo* sets the new few-shot state of the art. *Note:* We omit RareAct, our 16th benchmark, as it is a zero-shot benchmark with no available fine-tuned results to compare to. *Right*: Flamingo performance improves with model size and number of shots.

- 기존 모델과 달리 Flamingo는 고해상도 이미지와 비디오를 처리할 수 있다
- 텍스트 생성 작업을 위한 시각적 조건 부여를 효과적으로 수행
- 데이터가 부족한 상황에서도 탁월한 적응력

- 새로운 기준을 세우는 VLM으로, 캡셔닝, 시각적 대화, 시각적 질문-응답과 같은 개방형 작업을 단순히 몇 가지 입력/출력 예제로 적응할 수 있다.

⇒ 대규모 언어 모델의 설계 아이디어를 활용하여, 멀티모달 데이터를 효율적으로 처리하고 다양한 작업에 적용 가능한 범용성을 제공

- 기능
  - 텍스트와 이미지가 섞인 시퀀스를 다룰 수 있는 멀티모달 프롬프트 처리 능력
  - 고해상도 이미지 및 비디오를 처리할 수 있는 Perceiver 기반 아키텍처.
  - 사전 학습된 비전 모델과 언어 모델을 연결하는 새로운 아키텍처 구성 요소.

2. Approach

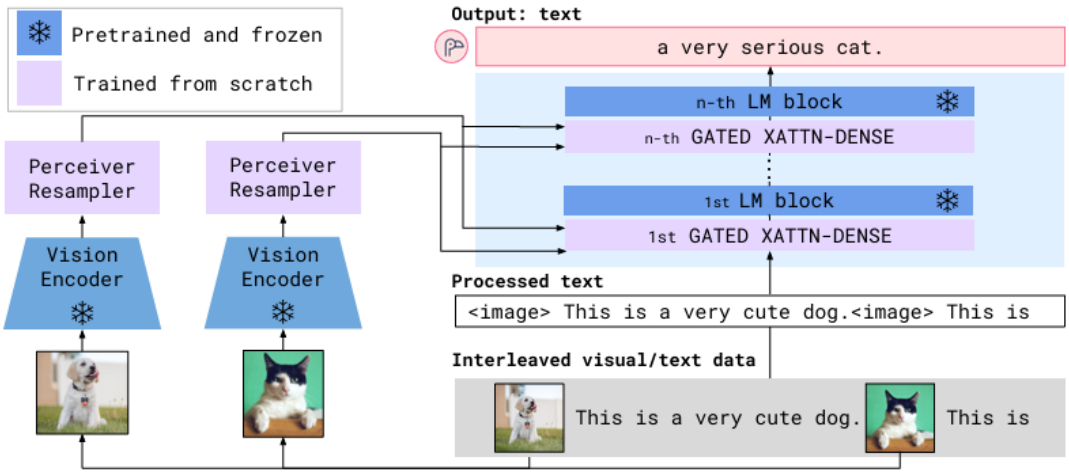

Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

텍스트와 이미지/비디오가 섞인 데이터를 입력으로 받아 자유 형식의 텍스트를 출력하는 비전-언어 모델

- 접근법
  - Perceive Resampler (2.1)

    : 비전 인코더에서 추출한 시각적 특징을 고정된 수의 시각적 토큰으로 변환하며, 이를 통해 시각적 데이터를 언어 모델에 통합
  - GATED XATTNDENSE (2.2)

: 시각적 정보를 사전 학습된 언어 모델에 효과적으로 연결하여, 다음 토큰 예측 과정에서 시각적 조건을 반영할 수 있게 한다.

$$\text{“}p(y \mid x) = \prod p(y_\ell \mid y < \ell, x \leq \ell),\text{”}$$

y_l : 입력 텍스트의 l번째 언어 토큰

$y_{<l}$: 선행 토큰의 집합

2.1 Visual processing and the Perceiver Resampler

Vision Encoder: from pixels to features.

고정된 Normalizer-Free ResNet(NFNet) 사용 : 대조적 목표를 사용하여 이미지와 텍스트 쌍 데이터셋에서 사전 학습

- 최종 단계에서 출력되는 2D 공간 격자 특징은 1D 시퀀스로 펼쳐진다.
- 비디오 입력의 경우, 초당 1프레임으로 샘플링된 프레임이 독립적으로 인코딩되어 3D 시공간 격자 특징을 생성한다. (시간 임베딩 추가)

Perceiver Resampler: from varying-size large feature maps to few visual tokens.

비전 인코더를 고정된 언어 모델과 연결하여, 생성된 이미지 또는 비디오 특징을 입력으로 받아, 64개의 시각적 출력을 생성하여 비전-텍스트 교차 주의 계산의 복잡성을 줄인다.

→ Perceiver Resampler는 Transformer와 교차 주의를 활용하여 다른 단순한 구조보다 우수한 성능을 보인다.

2.2 Conditioning frozen language models on visual representations

- **텍스트 생성은 Transformer Decoder 에 의해 수행되며**, Perceiver Resampler로부터 생성된 시각적 표현을 조건으로 한다.

Interleaving new GATED XATTN-DENSE layers within a frozen pretrained LM.

- 사전 학습된 고정된 텍스트 전용 언어 모델 레이어를, 새롭게 학습된 레이어와 교차 배치하여 시각적 출력을 Perceiver Resampler로부터 가져온다.

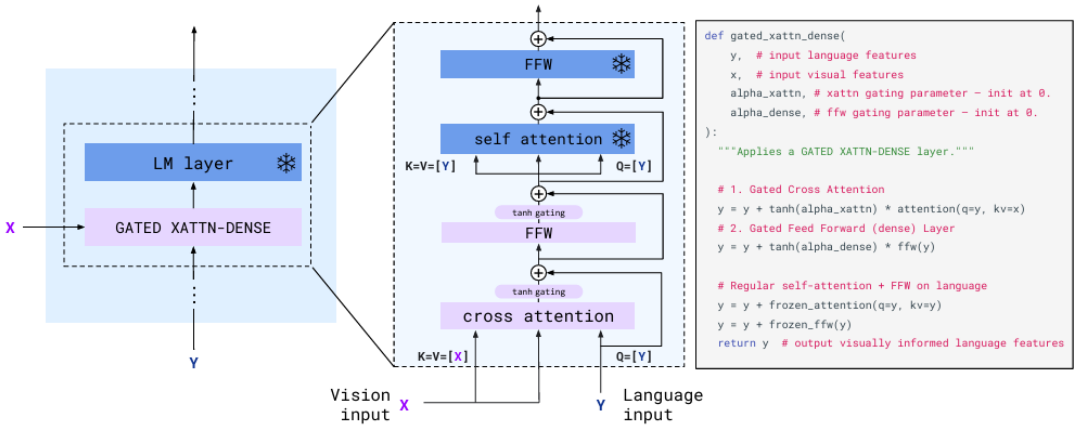

Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

⇒ 초기화 시점에서 조건화된 모델이 기존 언어 모델과 동일한 결과를 산출하도록 보장하여 학습 안정성과 최종 성능을 향상

Varying model sizes.
1.4B, 7B, 70B 파라미터를 가진 Chinchilla 모델에 기반하여 세 가지 크기로 실험

고정된 비전 인코더와 학습 가능한 Perceiver Resampler는 모델 크기와 관계없이 일정한 크기를 유지하며, 훈련 가능한 GATED XATTN-DENSE 모듈과 고정된 언어 모델의 파라미터 수는 증가한다.

2.3 Multi-visual input support: per-image/video attention masking

image-causal modeling은 the full text-to-image cross-attention matrix을 마스킹하여 얻어진다.

- 각 텍스트 토큰에서 모델이 텍스트 토큰 바로 앞에 나타난 이미지의 시각적 토큰만 참조하도록 제한
- 교차된 시퀀스의 이전 텍스트 토큰이나 모든 이전 이미지를 참조하는 대신, 해당 텍스트 바로 앞의 이미지와 관련된 시각적 토큰에만 attention을 두게 된다.

single-image cross-attention scheme은 모델이 훈련 시 사용된 이미지의 수와 관계없이 어떤 수의 시각적 입력에도 원활하게 일반화할 수 있도록 한다.

2.4 Training on a mixture of vision and language datasets

M3W:Interleaved image and text dataset.

- 약 4,300만 개의 웹페이지에서 HTML을 통해 텍스트와 이미지를 추출하고, 텍스트와 이미지 요소의 상대적 위치를 기반으로 텍스트와 이미지를 결합
- 각 문서에서 무작위로 256개의 토큰 시퀀스를 샘플링하며, 최대 5개의 이미지를 포함

Pairs of image/video and text.

- ALIGN 데이터셋(약 18억 개의 이미지와 alt-text로 구성)을 활용 + Long Text & Image Pairs(LTIP)라는 새로운 데이터셋(약 3억 1,200만 개의 이미지-텍스트 쌍)을 수집
- 비디오 입력을 다루기 위해, 약 2,700만 개의 짧은 비디오(평균 약 22초 길이)와 문장 설명으로 구성된 Video & Text Pairs(VTP) 데이터셋을 추가로 수집
- **모든 데이터셋은 M3W의 구문과 일치하도록 조정**

Multi-objective training and optimisation strategy.

$$\sum_{m=1}^{M} \lambda_m \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}_m} \left[-\sum_{\ell=1}^{L} \log p(y_\ell | y_{<\ell}, x_{\leq\ell}) \right],$$

m번째 데이터셋과 가중치를 의미한다.

2.5 Task adaptation with few-shot in-context learning

- **평가방법**
  - 개방형 평가 : beam search을 사용하여 디코딩을 수행
  - 폐쇄형 평가 : 모델의 로그 확률을 사용하여 각 가능한 답변을 스코어링

Flamingo는 지원 예제와 질의 시각적 입력을 교차 배치하는 방식으로 few-shot 학습을 수행. 이는 GPT-3와 유사한 방식으로 새로운 작업에 적응할 수 있는 능력을 제공

3. Experiments

3.1 Few-shot learning on vision-language tasks

Method	FT	Shot	OKVQA (I)	VQAv2 (I)	COCO (I)	MSVDQA (V)	VATEX (V)	VizWiz (I)	Flick30K (I)	MSRVTTQA (V)	iVQA (V)	YouCook2 (V)	STAR (V)	VisDial (I)	TextVQA (I)	NextQA (I)	HatefulMemes (I)	RareAct (V)
Zero/Few shot SOTA	✗	(X)	[34] 43.3 (16)	[114] 38.2 (4)	[124] 32.2 (0)	[58] 35.2 (0)	-	-	-	[58] 19.2 (0)	[135] 12.2 (0)	-	[143] 39.4 (0)	[79] 11.6 (0)	-	-	[85] 66.1 (0)	[85] 40.7 (0)
Flamingo-3B	✗	0	41.2	49.2	73.0	27.5	40.1	28.9	60.6	11.0	32.7	55.8	39.6	46.1	30.1	21.3	53.7	58.4
	✗	4	43.3	53.2	85.0	33.0	50.0	34.0	72.0	14.9	35.7	64.6	41.3	47.3	32.7	22.4	53.6	-
	✗	32	45.9	57.1	99.0	42.6	59.2	45.5	71.2	25.6	37.7	76.7	41.6	47.3	30.6	26.1	56.3	-
Flamingo-9B	✗	0	44.7	51.8	79.4	30.2	39.5	28.8	61.5	13.7	35.2	55.0	41.8	48.0	31.8	23.0	57.0	57.9
	✗	4	49.3	56.3	93.1	36.2	51.7	34.9	72.6	18.2	37.7	70.8	**42.8**	50.4	33.6	24.7	62.7	-
	✗	32	51.0	60.4	106.3	47.2	57.4	44.0	72.8	29.4	40.7	77.3	41.2	50.4	32.6	28.4	63.5	-
Flamingo	✗	0	50.6	56.3	84.3	35.6	46.7	31.6	67.2	17.4	40.7	60.1	39.7	52.0	35.0	26.7	46.4	**60.8**
	✗	4	57.4	63.1	103.2	41.7	56.0	39.6	75.1	23.9	44.1	74.5	42.4	**55.6**	36.5	30.8	68.6	-
	✗	32	**57.8**	**67.6**	**113.8**	**52.3**	**65.1**	**49.8**	**75.4**	**31.0**	**45.3**	**86.8**	42.2	**55.6**	**37.9**	**33.5**	**70.0**	-
Pretrained FT SOTA	✔	(X)	54.4 [34] (10K)	80.2 [140] (444K)	143.3 [124] (500K)	47.9 [28] (27K)	76.3 [153] (500K)	57.2 [65] (20K)	67.4 [150] (30K)	46.8 [51] (130K)	35.4 [135] (6K)	138.7 [132] (10K)	36.7 [128] (46K)	75.2 [79] (123K)	54.7 [137] (20K)	25.2 [129] (38K)	79.1 [62] (9K)	-

Table 1: **Comparison to the state of the art.** A *single* Flamingo model reaches the state of the art on a wide array of image **(I)** and video **(V)** understanding tasks with few-shot learning, significantly outperforming previous best zero- and few-shot methods with as few as four examples. More importantly, using only 32 examples and without adapting any model weights, Flamingo *outperforms* the current best methods – fine-tuned on thousands of annotated examples – on seven tasks. Best few-shot numbers are in **bold**, best numbers overall are underlined.

작업당 단 4개의 예제만으로도 이와 같은 성능을 달성하며, 시각적 모델을 새로운 작업에 실제적이고 효율적으로 적응하여 SOTA달성.

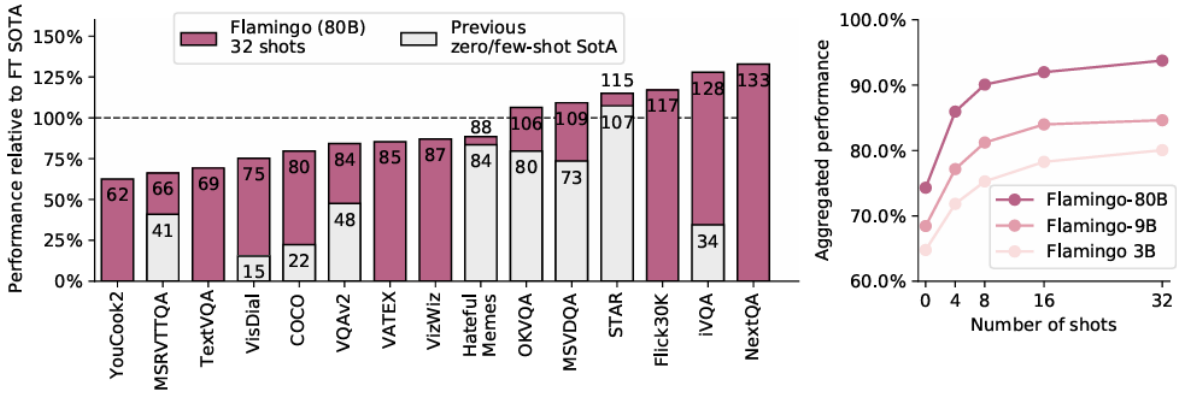

Figure 2: **Flamingo results overview.** *Left*: Our largest model, dubbed *Flamingo*, outperforms state-of-the-art fine-tuned models on 6 of the 16 tasks we consider with no fine-tuning. For the 9 tasks with published few-shot results, *Flamingo* sets the new few-shot state of the art. *Note:* We omit RareAct, our 16th benchmark, as it is a zero-shot benchmark with no available fine-tuned results to compare to. *Right*: Flamingo performance improves with model size and number of shots.

모델의 크기가 커질수록 성능이 더 좋아졌으며, 이는 GPT-3와 유사한 경향을 보였다.

3.2 Fine-tuning Flamingo as a pretrained vision-language model

목표: 더 많은 데이터를 제공받았을 때 Flamingo 모델이 특정 작업에 적응할 수 있는지 확인하기 위해 실험

Method	VQAV2		COCO	VATEX	VizWiz		MSRVTTQA	VisDial		YouCook2	TextVQA		HatefulMemes
	test-dev	test-std	test	test	test-dev	test-std	test	valid	test-std	valid	valid	test-std	test seen
32 shots	67.6	-	113.8	65.1	49.8	-	31.0	56.8	-	86.8	36.0	-	70.0
Fine-tuned	**82.0**	**82.1**	138.1	**84.2**	**65.7**	**65.4**	**47.4**	61.8	59.7	118.6	**57.1**	54.1	**86.6**
SotA	81.3† [133]	81.3† [133]	**149.6**† [119]	81.4† [153]	57.2† [65]	60.6† [65]	46.8 [51]	**75.2** [79]	**75.4**† [123]	**138.7** [132]	54.7 [137]	**73.7** [84]	84.6† [152]

Table 2: **Comparison to SotA when fine-tuning *Flamingo*.** We fine-tune *Flamingo* on all nine tasks where *Flamingo* does not achieve SotA with few-shot learning. *Flamingo* sets a new SotA on five of them, outperfoming methods (marked with †) that use tricks such as model ensembling or domain-specific metric optimisation (e.g., CIDEr optimisation).

- Flamingo는 제한 없는 주석 데이터를 사용하여 파인튜닝한 경우, few-shot learning 성능을 초과하는 결과를 보였다.
- Flamingo는 짧은 학습 스케줄과 낮은 학습률을 사용해 파인튜닝하며, 비전 백본을 추가로 해제하여 더 높은 해상도를 처리
-

3.3 Ablation studies

	Ablated setting	Flamingo 3B value	Changed value	Param. count ↓	Step time ↓	COCO CIDEr↑	OKVQA top1↑	VQAv2 top1↑	MSVDQA top1↑	VATEX CIDEr↑	Overall score↑
	Flamingo 3B model (short training)			3.2B	1.74s	86.5	42.1	55.8	36.3	53.4	**70.7**
(i)	Resampler size	Medium	Small	3.1B	1.58s	81.1	40.4	54.1	36.0	50.2	67.9
			Large	3.4B	1.87s	84.4	42.2	54.4	35.1	51.4	69.0
(ii)	Multi-Img att.	Only last	All previous	3.2B	1.74s	70.0	40.9	52.0	32.1	46.8	63.5
(iii)	p_{next}	0.5	0.0	3.2B	1.74s	85.0	41.6	55.2	36.7	50.6	69.6
			1.0	3.2B	1.74s	81.3	43.3	55.6	36.8	52.7	70.4
(iv)	LM pretraining	MassiveText	C4	3.2B	1.74s	81.3	34.4	47.1	60.6	53.9	62.8
(v)	Freezing Vision	✓	✗ (random init)	3.2B	4.70s*	74.5	41.6	52.7	31.4	35.8	61.4
			✗ (pretrained)	3.2B	4.70s*	83.5	40.6	55.1	34.6	50.7	68.1
(vi)	Co-train LM on MassiveText	✗	✓ (random init)	3.2B	5.34s*	69.3	29.9	46.1	28.1	45.5	55.9
			✓ (pretrained)	3.2B	5.34s*	83.0	42.5	53.3	35.1	51.1	68.6
(vii)	Dataset and Vision encoder	M3W+ITP+VTP and NFNetF6	LAION400M and CLIP	3.1B	0.86s	61.4	37.9	50.9	27.9	29.7	54.7
			M3W+LAION400M+VTP and CLIP	3.1B	1.58s	76.3	41.5	53.4	32.5	46.1	64.9

Table 10: **Additional ablation studies.** Each row in this ablation study table should be compared to the baseline Flamingo run reported at the top of the table. The step time measures the time spent to perform gradient updates on all training datasets. (*): Due to higher memory usage, these models were trained using four times more TPU chips. The obtained accumulation step time was therefore multiplied by four.

Ablation 연구는 Flamingo의 설계 결정이 성능에 미치는 영향을 체계적으로 평가하며, 데이터, 아키텍처, 학습 전략 모두 모델 성능 최적화에 중요한 요소임을 입증

4. Related Work

Language modelling and few-shot adaptation.

Language modelling은 Transformer의 도입 이후 상당한 발전을 이루었다. 대규모 데이터를 활용한 사전 학습 후 특정 task에 적응시키는 방식이 표준이 되었다. 이에 대한 많은 연구도 이루어졌다.

1. 어댑터 모듈 추가
2. Language model일부만 미세 조정
3. Prompt 내에서 예제 제공
4. Prompt 최적화
5. few-shot learning

When language meets vision

- BERT
- Flamingo는 새로운 작업에 대한 파인튜닝이 필요하지 않다.
- 대조적 학습을 기반으로 하는 모델

Web-scale vision and language training datasets.

수동으로 주석이 추가된 비전-언어 데이터셋은 구축 비용이 높아 크기가 제한적

→이를 해결하기 위한 방법: 자동으로 웹에서 수집된 비전-텍스트 데이터를 활용한 연구가 진행

Flamingo는 이러한 데이터에 추가로 텍스트와 이미지가 혼합된 웹 페이지를 단일 시퀀스로 훈련하는 방식의 중요성을 강조

5. Discussion

Limitations.

1. **언어 모델의 약점 상속**: Flamingo는 사전 학습된 언어 모델을 기반으로 하기 때문에, 언어 모델의 장기 시퀀스 처리 한계, 샘플 효율성 부족, 비논리적 출력을 그대로 갖습니다.
2. **분류 작업 성능 부족**: Flamingo는 개방형 작업에 강점을 가지지만, 텍스트-이미지 검색 및 분류 등의 특정 작업에서는 대조적 학습 기반 모델보다 성능이 낮습니다.
3. **in-context learning의 제약**: in-context learning은 소량의 데이터를 활용할 수 있어 효율적이지만, 더 많은 데이터가 제공되면 성능이 제한적이고 추론 비용이 높아질 수 있다.

Societal impacts.

- 긍정적 효과: Flamingo는 적은 데이터로도 고성능을 낼 수 있어 비전문가도 다양한 작업을 수행할 수 있다.
- 잠재적 위험: 성별 및 인종 편향, 불쾌한 출력, 사적 정보 누출 등의 위험이 언어 모델뿐만 아니라 시각적 입력으로 인해 더욱 확대될 가능성이 존재

Conclusion.

Flamingo는 다양한 이미지 및 비디오 작업에 적응할 수 있는 범용 멀티모달 모델로, 작업별 학

습 없이 뛰어난 성능을 제공한다. 또한 기존 비전-언어 작업의 범위를 넘어선 새로운 응용 가능성을 제시한다.